

Variational Approaches for Auto-Encoding Generative Adversarial Networks

Mihaela Rosca, Balaji Lakshminarayanan
David Warde-Farley, Shakir Mohamed

Presented by Shih-Ming Wang
ComputerVision Lab, UCSC

02-13-2019

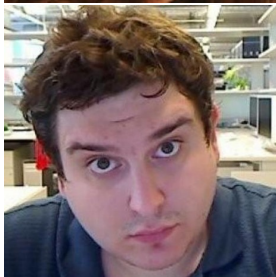
Authors



Mihaela Rosca
Research Engineer
DeepMind



Balaji Lakshminarayanan
Senior Research Scientist
DeepMind



David Warde-Farley
Senior Research Scientist
DeepMind



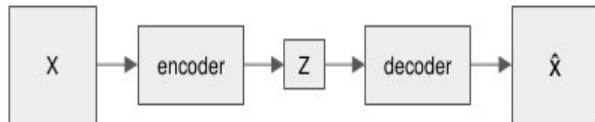
Shakir Mohamed
Research Scientist
DeepMind

Latent Variable Model I

- **Assumes** a generating process for real data $x \sim p^*(x)$
 - Unobserved quantity following prior distribution $z \sim p_\theta(z)$
 - Observation given z following likelihood $x|z \sim p_\theta(x|z)$
- Without losing generality, $z \sim \mathcal{N}(0, I)$ (dropping θ hereafter)
- Though $p_\theta(x, z) = p^*(x)p_\theta(z|x)$, it's not trivial to do inference (i.e. compute $p_\theta(z|x)$)
- Implicit Latent Variable Model, e.g. GAN
 - Learns a generator G_θ and makes likelihood implicit:

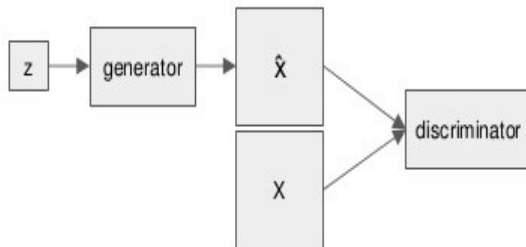
$$p_\theta(x|z) = \delta(x - G_\theta(z))$$
- Prescribed Latent Variable Model, e.g. VAE
 - **Assumes** explicit (closed form) likelihood $p_\theta(x|z)$
 - **Assumes** explicit (closed form) posterior $p_\theta(z|x)$
 - Approximates posterior $p_\theta(z|x)$, hence able to do inference

Latent Variable Model II



Variational
Autoencoders (VAE)

[Kingma and Welling
\[1312.6114\]](#)



Generative Adversarial
Networks (GAN)

[Goodfellow et al. \[1406.2661\]](#)

Motivation

Generative Adversarial Network (GAN) Pros & Cons

- Excels at generating sharp samples
- Only make assumption on prior
- Optimization is hard (vanishing gradient) using its original loss
- The samples might lack diversity (mode-collapse)
- Can't do inference

Variational Auto Encoder (VAE) Pros & Cons

- Generates blurry images
- Requires assumption on likelihood posterior, limiting model power
- Pairwise reconstruction penalty discourages mode-collapse
- Can do inference

KL divergence

- KL divergence measures “distance” between distributions

$$KL(p\|q) = \mathbb{E}_{p(x)} \left[\ln \frac{p(x)}{q(x)} \right] = \int p(x) \ln \frac{p(x)}{q(x)} dx$$

- $KL(p\|q) \geq 0$, with equality hold when $p(x) = q(x), \forall x$
- Requirement to approximate $KL(p\|q)$
 - Able to sample from $p(x)$
 - $p(x), q(x)$ bare close form
- In some cases, for e.g. when p, q are both Gaussian, $KL(p\|q)$ bares close form and sampling is not required
- KL is asymmetric, but Jensen-Shanon divergence (JSD) is

$$JS(p\|q) = 0.5KL(p\|(p+q)/2) + 0.5KL(q\|(p+q)/2)$$

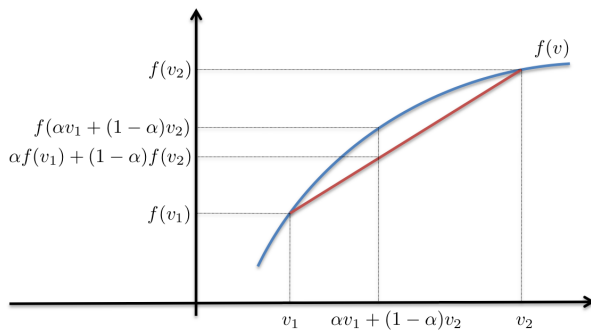
Jensen Inequality

- For concave function f and arbitrary distribution p

$$f(\mathbb{E}_{p(x)} [x]) \geq \mathbb{E}_{p(x)} [f(x)]$$

- Which implies for any function g

$$f(\mathbb{E}_{p(x)} [g(x)]) \geq \mathbb{E}_{p(x)} [f(g(x))]$$



Backpropagating through samples I

- Say we want to minimize $L(\theta, \eta) = \mathbb{E}_{p_\eta(x)} [f_\theta(x)]$
- Computing $\nabla_\theta L(\theta, \eta)$ is straightforward

$$\begin{aligned}\nabla_\theta L(\theta, \eta) &= \mathbb{E}_{p_\eta(x)} [\nabla_\theta L(\theta, \eta)] \\ &\approx \frac{1}{n} \sum_{i=1}^n \nabla_\theta f_\theta(x_i), \forall x_i \sim p_\eta(x)\end{aligned}$$

- However $\nabla_\eta L(\theta, \eta)$ is not in expectation form

$$\begin{aligned}\nabla_\eta L(\theta, \eta) &= \nabla_\eta \int p_\eta(x) f_\theta(x) dx \\ &= \int \nabla_\eta p_\eta(x) f_\theta(x) dx\end{aligned}$$

Backpropagating through samples II

Score function estimator (REINFORCE)

- use the identity $\nabla_{\eta} p_{\eta}(x) = p_{\eta}(x) \nabla_{\eta} \log p_{\eta}(x)$

$$\begin{aligned}\nabla_{\eta} L(\theta, \eta) &= \int \nabla_{\eta} p_{\eta}(x) f_{\theta}(x) dx \\ &= \int p_{\eta}(x) \nabla_{\eta} \log p_{\eta}(x) f_{\theta}(x) dx \\ &= \mathbb{E}_{p_{\eta}(x)} [f_{\theta}(x) \nabla_{\eta} \log p_{\eta}(x)] \\ &\approx \frac{1}{n} \sum_{i=1}^n f_{\theta}(x) \nabla_{\eta} \log p_{\eta}(x)\end{aligned}$$

- Such estimator of the gradient might have high variance

Backpropagating through samples III

Reparameterization Trick

- **Assume**

- $f_{\theta}(x)$ is differentiable w.r.t. x
- Exists g , $x = g(\eta, \epsilon)$, where $\epsilon \sim p(\epsilon)$.
- For e.g. $x \sim \mathcal{N}(\mu(\eta) + \sigma^2(\eta))$ can be reparameterize as $x = \mu(\eta) + \sigma^2(\eta) * \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$

- Then we get an gradient estimator with low variance

$$\begin{aligned}
 \nabla_{\eta} L(\theta, \eta) &= \nabla_{\eta} \mathbb{E}_{p_{\eta}(x)} [f_{\theta}(x)] \\
 &= \nabla_{\eta} \mathbb{E}_{p(\epsilon)} [f_{\theta}(g(\eta, \epsilon))] \\
 &= \mathbb{E}_{p(\epsilon)} [\nabla_{\eta} f_{\theta}(g(\eta, \epsilon))] \\
 &= \mathbb{E}_{p(\epsilon)} [f'_{\theta}(g(\eta, \epsilon)) \nabla_{\eta} g(\eta, \epsilon)]
 \end{aligned}$$

Maximum Log-Likelihood Principle (MLP)

- MLP optimizes θ so that $p_{\theta}(x) \rightarrow p^*(x)$

$$\theta = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^n \ln p_{\theta}(x_i), \forall x_i \sim p^*(x)$$

- MLP is equivalent to minimizing KL divergence

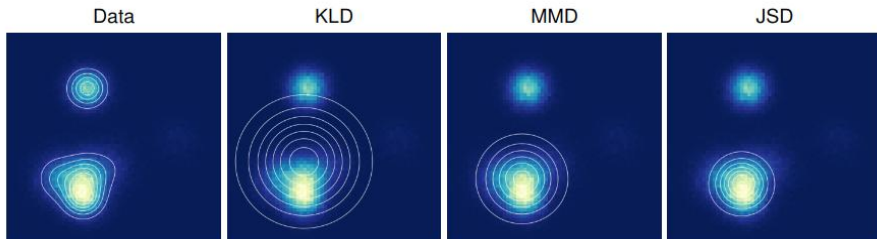
$$\begin{aligned} \underset{\theta}{\operatorname{argmin}} KL(p^*(x) \| p_{\theta}(x)) &= \underset{\theta}{\operatorname{argmin}} E_{p^*(x)} \left[\ln \frac{p^*(x)}{p_{\theta}(x)} \right] \\ &= \underset{\theta}{\operatorname{argmin}} E_{p^*(x)} [-\ln p_{\theta}(x)] + \underbrace{\mathbb{E}_{p^*(x)} [\ln p^*(x)]}_{\text{constant}} \\ &\sim \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^n \ln p_{\theta}(x_i), \forall x_i \sim p^*(x) \end{aligned}$$

- Marginal likelihood is intractable (no closed form) in Latent Variable Model

$$p_{\theta}(x) = \int p(z) p_{\theta}(x|z) dz$$

Choice of Loss

- When MLP is intractable, different training objective are adopted
- Most objectives are consistent with MLP given infinite data and model capacity
- With limited model capacity, different objective can lead to very different result
- GAN approximately minimizes JSD, leading to mode-collapse
- VAE approximates MLP (minimizes KLD), leading to unreal sample



Figures from [Theis et al., 2015]

GAN I

- GAN estimates $\frac{p^*(x)}{p_\theta(x)}$ with a discriminator D_ϕ
- The input to the discriminator follows $0.5p^*(x) + 0.5p_\theta(x)$

$$\frac{p^*(x)}{p_\theta(x)} = \frac{p_\phi(x|y=1)}{p_\phi(x|y=0)} = \frac{\cancel{p(x)}p_\phi(y=1|x)/\cancel{p(y=1)}}{\cancel{p(x)}p_\phi(y=0|x)/\cancel{p(y=0)}} \xrightarrow{0.5} \frac{D_\phi(x)}{1 - D_\phi(x)}$$

- The likelihood is $\prod_{i=1}^n D_\phi(x)^{y_i} (1 - D_\phi(x))^{1-y_i}$, hence MLP for ϕ is

$$\phi = \underset{\phi}{\operatorname{argmax}} \mathbb{E}_{p^*(x)} [\ln D_\phi(x)] + \mathbb{E}_{p_\theta(x)} [\ln(1 - D_\phi(x))]$$

- Fixing D_ϕ , optimizes the following loss w.r.t. the generator G_θ

$$\theta = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{p_\theta(x)} [\ln(1 - D_\phi(x))]$$

GAN II

- This is the minmax game that GAN played

$$\begin{aligned} & \min_{\theta} \max_{\phi} \mathbb{E}_{p^*(x)} [\ln D_{\phi}(x)] + \mathbb{E}_{p_{\theta}(x)} [\ln(1 - D_{\phi}(x))] \\ & \leftrightarrow \min_{\theta} \max_{\phi} \mathbb{E}_{p^*(x)} [\ln D_{\phi}(x)] + \mathbb{E}_{p(z)} [\ln(1 - D_{\phi}(G_{\theta}(z)))] \end{aligned}$$

- We need to be able to
 - sample from implicit likelihood $p_{\theta}(x|z)$
- We can do
 - sample from model distribution $p_{\theta}(x) = p(z)p_{\theta}(x|z)$
- We can't do
 - compute $p_{\theta}(x)$ given x
 - compute $p_{\theta}(z|x)$ or sample $z \sim p_{\theta}(z|x)$ given x

VAE I

- Variational Inference (VI) bounds $p_\theta(x)$ with evidence lower bound (ELBO) and follows MLE to maximize the ELBO:

$$\begin{aligned}
 \ln p_\theta(x) &= \ln \int p(z) p_\theta(x|z) dz \\
 &= \ln \int q_\eta(z) \frac{p(z)}{q_\eta(z)} p_\theta(x|z) dz \\
 &= \ln \mathbb{E}_{q_\eta(z)} \left[\frac{p(z)}{q_\eta(z)} p_\theta(x|z) \right] \\
 &\geq \mathbb{E}_{q_\eta(z)} \left[\ln \left(\frac{p(z)}{q_\eta(z)} p_\theta(x|z) \right) \right] && \text{(Jensen Inequality)} \\
 &= \mathbb{E}_{q_\eta(z)} \left[\ln \left(\frac{p_\theta(x, z)}{q_\eta(z)} \right) \right] && (ELBO_1) \\
 &= \mathbb{E}_{q_\eta(z)} [\ln p_\theta(x|z)] - KL(q_\eta(z) \| p(z)) && (ELBO_2)
 \end{aligned}$$

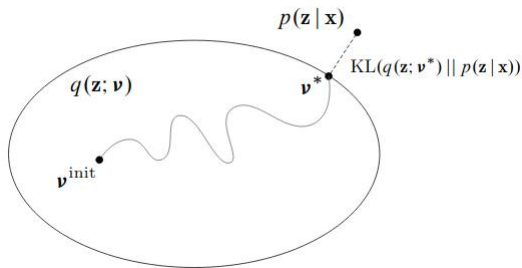
VAE II

- $q_\eta(z)$ (variational distribution) serves as approximate posterior
- EM algorithm alternates between the two steps
 - M-step: Fixing $q_\eta(z)$, optimize ELBO w.r.t. θ
 - E-step: Fixing p_θ , optimize ELBO w.r.t. $q_\eta(z)$

$$\begin{aligned}
 q_\eta(z) &= \underset{q_\eta(z)}{\operatorname{argmax}} \mathbb{E}_{q_\eta(z)} \left[\ln \left(\frac{p_\theta(x, z)}{q_\eta(z)} \right) \right] && (ELBO_1) \\
 &= \underset{q_\eta(z)}{\operatorname{argmax}} \mathbb{E}_{q_\eta(z)} \left[\ln \left(\frac{p(x)p_\theta(z|x)}{q_\eta(z)} \right) \right] \\
 &= \underset{q_\eta(z)}{\operatorname{argmax}} \ln p(x) + \mathbb{E}_{q_\eta(z)} \left[\ln \left(\frac{p_\theta(z|x)}{q_\eta(z)} \right) \right] \\
 &= \underset{q_\eta(z)}{\operatorname{argmax}} \ln p(x) - KL(q_\eta(z) \| p_\theta(z|x)) \\
 &= p_\theta(z|x)
 \end{aligned}$$

VAE III

- $p_\theta(x)$ in $p_\theta(z|x) = \frac{p(z)p_\theta(x|z)}{p(x)}$ is usually intractable
- Usually assume parametric form $q_\eta(z)$ and optimize η so that $q_\eta(z) \rightarrow p_\theta(z|x)$
- $q_\eta(z)$ usually has local parameters for each sample x . Alternatively, we can “armortize” it by modeling $q_\eta(z|x)$ with global parameters



Figures from [Blei et al.,]

VAE IV

- VAE models $q_\eta(z|x)$ with encoder E_η and $p_\theta(x|z)$ with decoder G_θ
- **Assumes** posterior $q_\eta(z|x_n) \sim \mathcal{N}(E_\eta(x_n), E_\eta(x_n)\mathbb{I})$
- **Assumes** likelihood $p_\theta(x_n|z) \sim \mathcal{N}(G_\theta(x_n), \mathbb{I})$
- Train η, θ simultaneously to minimize $ELBO_2$

$$\eta, \theta = \underset{\eta, \theta}{\operatorname{argmax}} \mathbb{E}_{q_\eta(z|x)} [\ln p_\theta(x|z)] - KL(q_\eta(z|x) \| p(z))$$

- VAE use reparameterization trick to estimate $\nabla_\eta \mathbb{E}_{q_\eta(z|x)} [\ln p_\theta(x|z)]$
- The first term is the reconstruction loss (auto-encoder)
- The second term serves a regularization, leading to higher reconstruction error

VAE V

- This is the game that VAE played

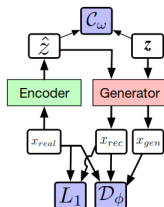
$$\eta, \theta = \underset{\eta, \theta}{\operatorname{argmax}} \mathbb{E}_{q_{\eta}(z|x)} [\ln p_{\theta}(x|z)] - KL(q_{\eta}(z|x) \| p(z))$$

- We need to be able to
 - compute likelihood $p_{\theta}(x|z)$
 - compute posterior approximation $q_{\eta}(z|x)$
- We can do
 - sample from model distribution $p_{\theta}(x) = p(z)p_{\theta}(x|z)$
 - compute $q_{\theta}(z|x)$ or sample $z \sim q_{\theta}(z|x)$ given x
- We can't do
 - compute $p_{\theta}(x)$ given x

Fusion of VAE & GAN

Key Idea

- Starts with VAE architecture
- Remove the close form assumption for $q_\eta(z|x)$ and $p_\theta(x|z)$
- Use implicit distribution for
 - posterior: $q_\eta(z|x) = \delta(z - E_\eta(x))$
 - likelihood: $p_\theta(x|z) = \delta(x - G_\theta(z))$
- Use the GAN trick to learn two discriminators C_ω and D_ϕ to estimate the two terms in ELBO



Implicit Variational Decomposition

- Instead of Gaussian, let $q_\theta(z|x) = \delta(z - E_\eta(x))$
- The KL term in ELBO can be estimated by

$$KL(q_\eta(z|x)||p(z)) = \mathbb{E}_{q_\eta(z|x)} \left[\ln \frac{q_\eta(z|x)}{p(z)} \right] \approx \mathbb{E}_{q_\eta(z|x)} \left[\ln \frac{C_\omega(z)}{1 - C_\omega(z)} \right]$$

- C_ω is a discriminator trained to tell samples drawn from $q_\eta(z|x)$ or $p(z)$

Hybrid Likelihood

- Let $p_\theta(x|z) = \delta(x - G_\theta(z))$
- The reconstruction term in ELBO can be estimated by

$$\begin{aligned}\mathbb{E}_{q_\eta(z|x)} [\ln p_\theta(x|z)] &= \mathbb{E}_{q_\eta(z|x)} \left[\ln \left(\frac{p_\theta(x|z)}{p^*(x)} p^*(x) \right) \right] \\ &= \mathbb{E}_{q_\eta(z|x)} \left[\ln \frac{p_\theta(x|z)}{p^*(x)} \right] + \cancel{\ln p^*(x)} \\ &\approx \mathbb{E}_{q_\eta(z|x)} \left[\ln \frac{D_\phi(G_\theta(z))}{1 - D_\phi(G_\theta(z))} \right]\end{aligned}$$

- D_ϕ is a discriminator trained to tell samples from $p_\theta(x|z)$ or $p^*(x)$
- We can hybrid the implicit likelihood with an explicit likelihood
- For e.g. Laplace: $p_\theta(x|z) \propto \exp(-\lambda \|x - G_\theta(z)\|_1)$, which leads to the L_1 reconstruction loss $\mathbb{E}_{q_\eta(z|x)} [-\lambda \|x - G_\theta(z)\|_1]$

α -GAN I

- This is the total loss of α - GAN

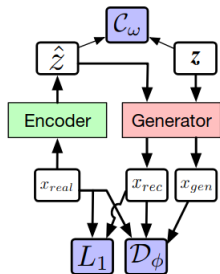
$$\mathcal{L}(\theta, \eta) = \mathbb{E}_{q_{\eta}(z|x)} \left[-\lambda \|x - G_{\theta}(z)\|_1 + \ln \frac{D_{\phi}(G_{\theta}(z))}{1 - D_{\phi}(G_{\theta}(z))} + \ln \frac{C_{\omega}(z)}{1 - C_{\omega}(z)} \right]$$

- The loss for D_{ϕ} and C_{ω} are not shown
- We need to be able to
 - train encoder E_{η} , decoder G_{θ} , and discriminators D_{ϕ} , C_{ω}
 - sample from implicit posterior $z \sim q_{\eta}(z|x)$
 - sample from implicit likelihood $x \sim p_{\theta}(x|z)$
 - compute explicit likelihood $p_{\theta}(x|z)$
- We can do
 - sample from model distribution $p_{\theta}(x) = p(z)p_{\theta}(x|z)$
 - sample from implicit posterior $q_{\theta}(z|x)$ given x
- We can't do
 - compute $p_{\theta}(x)$ given x
 - compute $q_{\theta}(z|x)$ given x

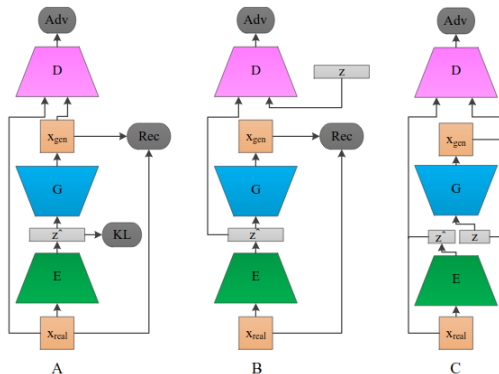
α -GAN II

Tricks to improve

- For the two discriminator, use a different loss to avoid vanishing gradient
- Pass reconstruct sample $x_{rec} \sim G_{\theta}(E_{\eta}(x))$ as well as sample from noise $x_{gen} \sim \sim G_{\theta}(z)$ to train D_{ϕ}



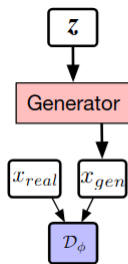
Related Work I



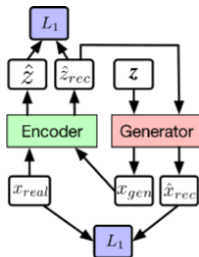
Figures from [Huang et al., 2018]

- Ⓐ VAEGAN imposes a discriminator (D_ϕ) on the data space
- Ⓑ AAE imposes a discriminator (C_ω) on the latent space
- Ⓒ ALI & BiGan discriminate **jointly** in the data and latent space

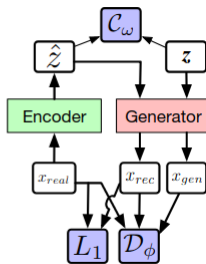
Related Work II



(a) DCGAN



(b) AGE

(c) α -GAN

- DCGAN is the normal GAN with some architecture tweaks
- AGE adds a **loop** from Generator back to Encoder, which is used as discriminator
- α – GAN imposes two discriminator on the data space and the latent space

Evaluation

- Inception Score: samples should cluster compactly

$$\mathbb{E}_x [KL(p(y|x)||p(y))]$$

- (1 - avg pairwise MS-SSIM): test in-class mode collapse (CelebA only)
- Wasserstein critic: discriminator to tell samples from train set or val set. Capture memorization and mode-collapse

Experiments

- Compared with DC-GAN, WGAN-GP, AGE
- Using ColorMnist, CelebA, CIFAR10 dataset
- Observations:
 - WGAN-GP, having no inference power, wins α - GAN most of the time
 - α - GAN wins AGE, having inference power, most of the time
 - By visual inspection, WGAN-GP still generates more realistic samples.
It's hard to tell α - GAN generates better sample than AGE

 Blei, D., Ranganath, R., and Mohamed, S.

Variational inference: Foundations and modern methods.

 Huang, H., He, R., Sun, Z., Tan, T., et al. (2018).

Introvae: Introspective variational autoencoders for photographic image synthesis.

In Advances in Neural Information Processing Systems, pages 52–63.

 Theis, L., Oord, A. v. d., and Bethge, M. (2015).

A note on the evaluation of generative models.

arXiv preprint arXiv:1511.01844.